Exploiting Image Generality for Lexical Entailment Detection

Douwe Kiela
Computer Laboratory
University of Cambridge
douwe.kiela@cl.cam.ac.uk

Laura Rimell
Computer Laboratory
University of Cambridge
laura.rimell@cl.cam.ac.uk

Ivan Vulić
Department of Computer Science
KU Leuven
ivan.vulic@cs.kuleuven.be

Stephen Clark
Computer Laboratory
University of Cambridge
stephen.clark@cl.cam.ac.uk

Abstract

We exploit the visual properties of concepts for lexical entailment detection by examining a concept's *generality*. We introduce three unsupervised methods for determining a concept's generality, based on its related images, and obtain state-of-the-art performance on two standard semantic evaluation datasets. We also introduce a novel task that combines hypernym detection and directionality, significantly outperforming a competitive frequency-based baseline.

1 Introduction

Automatic detection of lexical entailment is useful for a number of NLP tasks including search query expansion (Shekarpour et al., 2013), recognising textual entailment (Garrette et al., 2011), metaphor detection (Mohler et al., 2013), and text generation (Biran and McKeown, 2013). Given two semantically related words, a key aspect of detecting lexical entailment, or the hyponym-hypernym relation, is the *generality* of the hypernym compared to the hyponym. For example, *bird* is more general than *eagle*, having a broader intension and a larger extension. This property has led to the introduction of lexical entailment measures that compare the entropy of distributional word representations, under the assumption that a more general term has a higher-entropy distribution (Herbelot and Ganesalingam, 2013; Santus et al., 2014).

A strand of distributional semantics has recently emerged that exploits the fact that meaning is often grounded in the perceptual system, known as multi-modal distributional semantics (Bruni et al., 2014). Such models enhance purely linguistic models with extra-linguistic perceptual information, and outperform language-only models on a range of tasks, including modelling semantic similarity and conceptual relatedness (Silberer and Lapata, 2014). In fact, under some conditions uni-modal visual representations outperform traditional linguistic representations on semantic tasks (Kiela and Bottou, 2014).

We hypothesize that visual representations can be particularly useful for lexical entailment detection. Deselaers and Ferrari (2011) have shown that sets of images corresponding to terms at higher levels in the WordNet hierarchy have greater visual variability than those at lower levels. We exploit this tendency using sets of images returned by Google's image search. The intuition is that the set of images returned for *animal* will consist of pictures of different kinds of animals, the set of images for *bird* will consist of pictures of different birds, while the set for *owl* will mostly consist only of images of owls, as can be seen in Figure 1.

Here we evaluate three different vision-based methods for measuring term generality on the semantic tasks of hypernym detection and hypernym directionality. Using this simple yet effective unsupervised approach, we obtain state-of-the-art results compared with supervised algorithms which use linguistic data.

2 Related Work

In the linguistic modality, the most closely related work is by Herbelot and Ganesalingam (2013) and Santus et al. (2014), who use unsupervised distributional generality measures to identify the hypernym in a hyponym-hypernym pair. Herbelot and Ganesalingam (2013) use KL divergence to compare the probability distribution of context words, given a term, to the background probability distribution of context words. Santus et al. (2014) use the median entropy of the probability distributions associated with a term's top-weighted con-

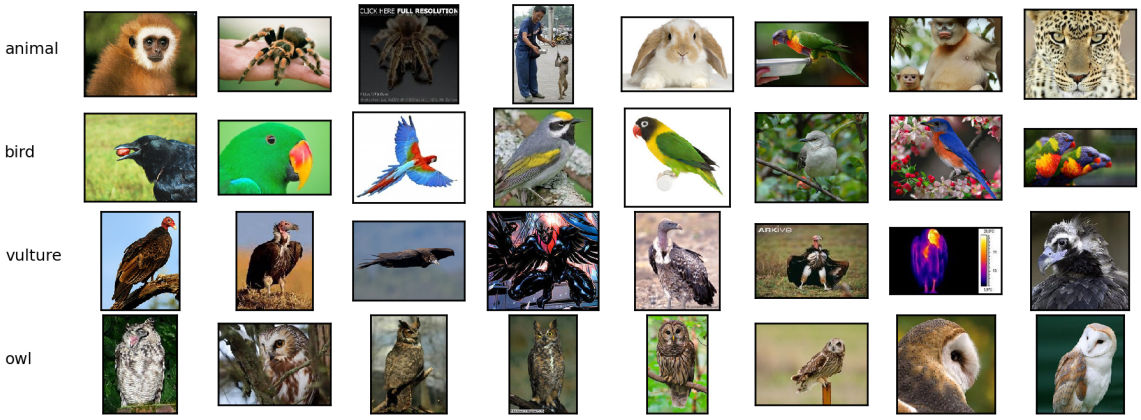

Figure 1: Example of how *vulture* and *owl* are less dispersed concepts than *bird* and *animal*, according to images returned by Google image search.

text words as a measure of information content.

In the visual modality, the intuition that visual representations may be useful for detecting lexical entailment is inspired by Deselaers and Ferrari (2011). Using manually annotated images from ImageNet (Deng et al., 2009), they find that concepts and categories with narrower intensions and smaller extensions tend to have less visual variability. We extend this intuition to the unsupervised setting of Google image search results and apply it to the lexical entailment task.

3 Approach

We use two standard evaluations for lexical entailment: hypernym directionality, where the task is to predict which of two words is the hypernym; and hypernym detection, where the task is to predict whether two words are in a hypernym-hyponym relation (Weeds et al., 2014; Santus et al., 2014). We also introduce a third, more challenging, evaluation that combines detection and directionality.

For the directionality experiment, we evaluate on the hypernym subset of the well-known BLESS dataset (Baroni and Lenci, 2011), which consists of 1337 hyponym-hypernym pairs. In this case, it is known that the words are in an entailment relation and the task is to predict the directionality of the relation. BLESS data is always presented with the hyponym first, so we report how often our measures predict that the second term in the pair is more general than the first.

For the detection experiment, we evaluate on the BLESS-based dataset of Weeds et al. (2014), which consists of 1168 word pairs and which we call WBLESS. In this dataset, the positive examples are hyponym-hypernym pairs. The negative examples include pairs in the reversed hypernym-hyponym order, as well as holonym-meronym pairs, co-hyponyms, and randomly matched nouns. Accuracy on WBLESS reflects the ability to distinguish hypernymy from other relations, but does not require detection of directionality, since reversed pairs are grouped with the other negatives.

BLESS	turtle—animal	1
WBLESS	owl—creature	1
	owl—vulture	0
	animal—owl	0
BIBLESS	owl—creature	1
	owl—vulture	0
	animal—owl	-1

Table 1: Examples for evaluation datasets.

For the combined experiment, we assign reversed hyponym-hypernym pairs a value of -1 instead of 0. We call this more challenging dataset BIBLESS. Examples of pairs in the respective datasets can be found in Table 1.

3.1 Image representations

Following previous work in multi-modal semantics (Bergsma and Goebel, 2011; Kiela et al., 2014), we obtain images from *Google Images*[1] for the words in the evaluation datasets. It has been shown that images from Google yield higher-quality representations than comparable resources such as Flickr and are competitive with "hand prepared datasets" (Bergsma and Goebel, 2011; Fergus et al., 2005).

[1] `www.google.com/imghp`. Images were retrieved on 10 April, 2015 from Cambridge in the United Kingdom.

For each image, we extract the pre-softmax layer from a forward pass in a convolutional neural network (CNN) that has been trained on the ImageNet classification task using Caffe (Jia et al., 2014). As such, this work is an instance of deep transfer learning; that is, a deep learning representation trained on one task (image classification) is used to make predictions on a different task (image generality). We chose to use CNN-derived image representations because they have been found to be of higher quality than the traditional bag of visual words models (Sivic and Zisserman, 2003) that have previously been used in multi-modal distributional semantics (Bruni et al., 2014; Kiela and Bottou, 2014).

3.2 Generality measures

We propose three measures that can be used to calculate the generality of a set of images. The image *dispersion* d of a concept word w is defined as the average pairwise cosine distance between all image representations $\{\vec{w_1}\ ...\ \vec{w_n}\}$ of the set of images returned for w:

$$d(w) = \frac{2}{n(n-1)} \sum_{i<j\leq n} 1 - cos(\vec{w_i}, \vec{w_j}) \quad (1)$$

This measure was originally introduced to account for the fact that perceptual information is more relevant for e.g. *elephant* than it is for *happiness*. It acts as a substitute for the concreteness of a word and can be used to regulate how much perceptual information should be included in a multi-modal model (Kiela et al., 2014).

Our second measure follows Deselaers and Ferrari (2011), who take a similar approach but instead of calculating the pairwise distance calculate the distance to the *centroid* $\vec{\mu}$ of $\{\vec{w_1}\ ...\ \vec{w_n}\}$:

$$c(w) = \frac{1}{n} \sum_{1\leq i\leq n} 1 - cos(\vec{w_i}, \vec{\mu}) \quad (2)$$

For our third measure we follow Lazaridou et al. (2015), who try different ways of modulating the inclusion of perceptual input in their multi-modal skip-gram model, and find that the *entropy* of the centroid vector $\vec{\mu}$ works well (where $p(\mu_j) = \frac{\mu_j}{||\vec{\mu}||}$ and m is the vector length):

$$H(w) = -\sum_{j=1}^{m} p(\mu_j) \log_2(p(\mu_j)) \quad (3)$$

3.3 Hypernym Detection and Directionality

We calculate the directionality of a hyponym-hypernym pair with a measure f using the following formula for a word pair (p, q). Since even co-hyponyms will not have identical values for f, we introduce a threshold α which sets a minimum difference in generality for hypernym identification:

$$s(p, q) = 1 - \frac{f(p) + \alpha}{f(q)} \quad (4)$$

In other words, $s(p, q) > 0$ iff $f(q) > f(p) + \alpha$, i.e. if the second word (q) is (sufficiently) more general. To avoid false positives where one word is more general but the pair is not semantically related, we introduce a second threshold θ which sets f to zero if the two concepts have low cosine similarity. This leads to the following formula:

$$s_\theta(p, q) = \begin{cases} 1 - \frac{f(p)+\alpha}{f(q)} & \text{if } \cos(\vec{\mu_p}, \vec{\mu_q}) \geq \theta \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

We experimented with different methods for obtaining the mean vector representations μ_c in Equation (5) and found that multi-modal representations worked best, concatenating the L2-normalized visual and linguistic representations to obtain a multi-modal representation, following Kiela and Bottou (2014). In other words, $\mu_c = w^{ling} \mid\mid \frac{1}{n}\sum_i^n w_i^{img}$. For comparison, we also report results for a visual-only μ_c.

For BLESS, we know the words in a pair stand in an entailment relation, so we set $\alpha = \theta = 0$ and evaluate whether $s(p, q) > 0$, indicating that q is a hypernym of p. For WBLESS, we set $\alpha = 0.02$ and $\theta = 0.2$ without tuning, and evaluate whether $s_\theta(p, q) > 0$ (hypernym relation) or $s_\theta(p, q) \leq 0$ (no hypernym relation). For BIBLESS, we set $\alpha = 0.02$ and $\theta = 0.25$ without tuning, and evaluate whether $s_\theta(p, q) > 0$ (hyponym-hypernym), $s(p, q) = 0$ (no relation), or $s(p, q) \leq 0$ (hypernym-hyponym).

4 Results

The results can be found in Table 2. We compare our methods with a frequency baseline, setting $f(p) = \text{freq}(p)$ in Equation 4 and using the frequency scores from Turney et al. (2011). Frequency has been proven to be a surprisingly challenging baseline for hypernym directionality (Herbelot and Ganesalingam, 2013; Weeds et al.,

	BLESS	WBLESS	BIBLESS
Frequency	0.58	0.57	0.39
WeedsPrec	0.63	—	—
WeedsSVM	—	0.75	—
WeedsUnSup	—	0.58	—
SLQS	0.87	—	—
Dispersion	0.88	0.75 (0.74)	0.57 (0.55)
Centroid	0.87	0.74 (0.74)	0.57 (0.54)
Entropy	0.83	0.71 (0.71)	0.56 (0.53)

Table 2: Accuracy. For WBLESS and BIBLESS we report results for multi-modal μ_c, with visual-only μ_c in brackets.

2014). In addition, we compare to the reported results of Santus et al. (2014) for WeedsPrec (Weeds et al., 2004), an early lexical entailment measure, and SLQS, the entropy-based method of Santus et al. (2014). Note, however, that these are on a subsampled corpus of 1277 word pairs from BLESS, so the results are indicative but not directly comparable. On WBLESS we compare to the reported results of Weeds et al. (2014): we include results for the highest-performing supervised method (WeedsSVM) and the highest-performing unsupervised method (WeedsUnSup).

For BLESS, both dispersion and centroid distance reach or outperform the best other measure (SLQS). They beat the frequency baseline by a large margin (+30% and +29%). Taking the entropy of the mean image representations does not appear to do as well as the other two methods but still outperforms the baseline and WeedsPrec (+25% and +20% respectively).

In the case of WBLESS and BIBLESS, we see a similar pattern in that dispersion and centroid distance perform best. For WBLESS, these methods outperform the other unsupervised approach, WeedsUnsup, by +17% and match the best-performing support vector machine (SVM) approach in Weeds et al. (2014). In fact, Weeds et al. (2014) report results for a total of 6 supervised methods (based on SVM and k-nearest neighbor (k-NN) classifiers): our unsupervised image dispersion method outperforms all of these except for the highest-performing one, reported here.

We can see that the task becomes increasingly difficult as we go from directionality to detection to the combination: the dispersion-based method goes from 0.88 to 0.75 to 0.57, for example. BIBLESS is the most difficult, as shown by the frequency baseline obtaining only 0.39. Our methods do much better than this baseline (+18%). Image dispersion appears to be the most robust measure.

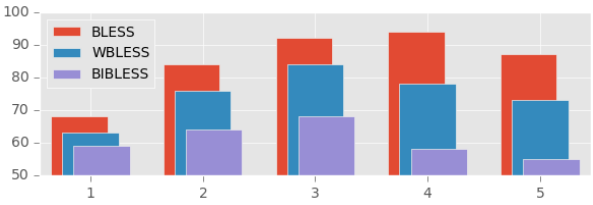

Figure 2: Accuracy by WordNet shortest path bucket (1 is shortest, 5 is longest).

To examine our results further, we divided the test data into buckets by the shortest WordNet path connecting word pairs (Miller, 1995). We expect our method to be less accurate on word pairs with short paths, since the difference in generality may be difficult to discern. It has also been suggested that very abstract hypernyms such as *object* and *entity* are difficult to detect because their linguistic distributions are not supersets of their hyponyms' distributions (Rimell, 2014), a factor that should not affect the visual modality. We find that concept comparisons with a very short path (bucket 1) are indeed the least accurate. We also find some drop in accuracy on the longest paths (bucket 5), especially for WBLESS and BIBLESS, perhaps because semantic similarity is difficult to detect in these cases. For a histogram of the accuracy scores according to WordNet similarity, see Figure 2.

5 Conclusions

We have evaluated three unsupervised methods for determining the generality of a concept based on its visual properties. Our best-performing method, image dispersion, reaches the state-of-the-art on two standard semantic evaluation datasets. We introduced a novel, more difficult task combining hypernym detection and directionality, and showed that our methods outperform a frequency baseline by a large margin.

We believe that image generality may be particularly suited to entailment detection because it does not suffer from the same issues as linguistic distributional generality. Herbelot and Ganesalingam (2013) found that general terms like *liquid* do not always have higher entropy distributions than their hyponyms, since speakers use them in very specific contexts, e.g. *liquid* is often coordinated with *gas*.

We also acknowledge that our method depends to some degree on Google's search result diversification, but do not feel this detracts from the utility

of the method, since the fact that general concepts achieve greater maximum image dispersion than specific concepts is not dependent on any particular diversification algorithm. In future work, we plan to explore more sophisticated visual generality measures, other semantic relations and different ways of fusing visual representations with linguistic knowledge.

Acknowledgments

DK and LR are supported by EPSRC grant EP/I037512/1. IV is supported by the PARIS project (IWT-SBO 110067) and the PDM Kort postdoctoral fellowship from KU Leuven. SC is supported by ERC Starting Grant DisCoTex (306920) and EPSRC grant EP/I037512/1. We thank the anonymous reviewers for their helpful comments.

References

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop*, pages 1–10.

Shane Bergsma and Randy Goebel. 2011. Using visual information to predict lexical preference. In *Proceedings of RANLP*, pages 399–405.

Or Biran and Kathleen McKeown. 2013. Classifying taxonomic relations between pairs of wikipedia articles. In *Proceedings of IJCNLP*, pages 788–794.

Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference. In *Proceedings of EMNLP*, pages 628–635.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Daoud Clarke. 2009. Context-theoretic semantics for natural language: An overview. In *Proceedings of the GEMS 2009 Workshop*, pages 112–119.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of CVPR*, pages 248–255.

Thomas Deselaers and Vittorio Ferrari. 2011. Visual and semantic similarity in imagenet. In *Proceedings of CVPR*, pages 1777–1784.

Robert Fergus, Li Fei-Fei, Pietro Perona, and Andrew Zisserman. 2005. Learning object categories from Google's image search. In *Proceedings of ICCV*, pages 1816–1823.

Dan Garrette, Katrin Erk, and Raymond Mooney. 2011. Integrating logical representations with probabilistic information using Markov logic. In *Proceedings of IWCS*, pages 105–114.

M. Geffet and I. Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of ACL*, pages 107–114.

Aurélie Herbelot and Mohan Ganesalingam. 2013. Measuring semantic content in distributional vectors. In *Proceedings of ACL*, pages 440–445.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Hongyan Jing. 1998. Usage of wordnet in natural language generation. In *Proceedings of COLING-ACL'98 Workshop on Usage of WordNet in Natural Language Processing Systems*.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of EMNLP*, pages 36–45.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of ACL*, pages 835–841.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of NAACL-HLT*.

Alessandro Lenci and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *Proceedings of *SEM*, pages 75–79.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of NAACL*.

J.H. Martin. 1990. *A Computational Model of Metaphor Interpretation*. Academic Press Professional, Inc.

George A. Miller. 1995. WordNet: A lexical database for English. In *Communications of the ACM*, volume 38, pages 39–41.

Michael Mohler, David Bracewell, Marc Tomlinson, and David Hinote. 2013. Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the 1st Workshop on Metaphor in NLP*.

Laura Rimell. 2014. Distributional lexical entailment by topic coherence. In *Proceedings of EACL*, pages 511–519.

Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte im Walde. 2014. Chasing hypernyms in vector spaces with entropy. In *Proceedings of EACL*, pages 38–42.

Saeedeh Shekarpour, Konrad Höffner, Jens Lehmann, and Sören Auer. 2013. Keyword query expansion on linked data using linguistic and semantic features. In *Proceedings of the 7th IEEE International Conference on Semantic Computing*, pages 191–197.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of ACL*, pages 721–732.

Josef Sivic and Andrew Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of ICCV*, pages 1470–1477.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of EMNLP*, pages 680–690.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of COLING*.

Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. 2014. Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of COLING*, pages 2249–2259.

W. A. Woods, Stephen Green, Paul Martin, and Ann Houston. 2001. Aggressive morphology and lexical relations for query expansion. In *Proceedings of TREC*.

M. Zhitomirsky-Geffet and I. Dagan. 2009. Bootstrapping distributional feature vector quality. *Computational Linguistics*, 35(3):435461.